Introduction to Statistical Methods in Political Science

Lecture 8: Statistical Inference for One Proportion

Ignacio Urbina

Ph.D. Candidate in Political Science

Learning Objectives

- **Understand:**
  - Confidence Intervals for p
  - Hypothesis Testing for p

Sample Proportion - Sampling Distribution

Inference with Proportions

- **Key Terms:**
  - **Population Proportion (p):**

$$p = \frac{\text{Number of successes in population}}{\text{Total population size}} = \frac{S}{N}$$

  - **Sample Proportion ($\hat{p}$):**

$$\hat{p} = \frac{S_n}{n}$$

$$S_n = \text{Number of successes in the sample}, \quad n = \text{Sample size}$$

- **Purpose:** Infer p from $\hat{p}$

Intuition Behind Inference

- **Concept:** Statistical inference involves estimating an unknown population parameter based on sample data.
- **Sample Variability:** Measuring the same quantity multiple times with slight variations each time.

$$\text{Repeated sampling} \Rightarrow \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$$

- **Example**: Assume you collect data from a sample of size n, and you compute the proportion of people vaccinated against some virus, $\hat{p}$.

Visualizing Sample Variability in Vaccination Rates

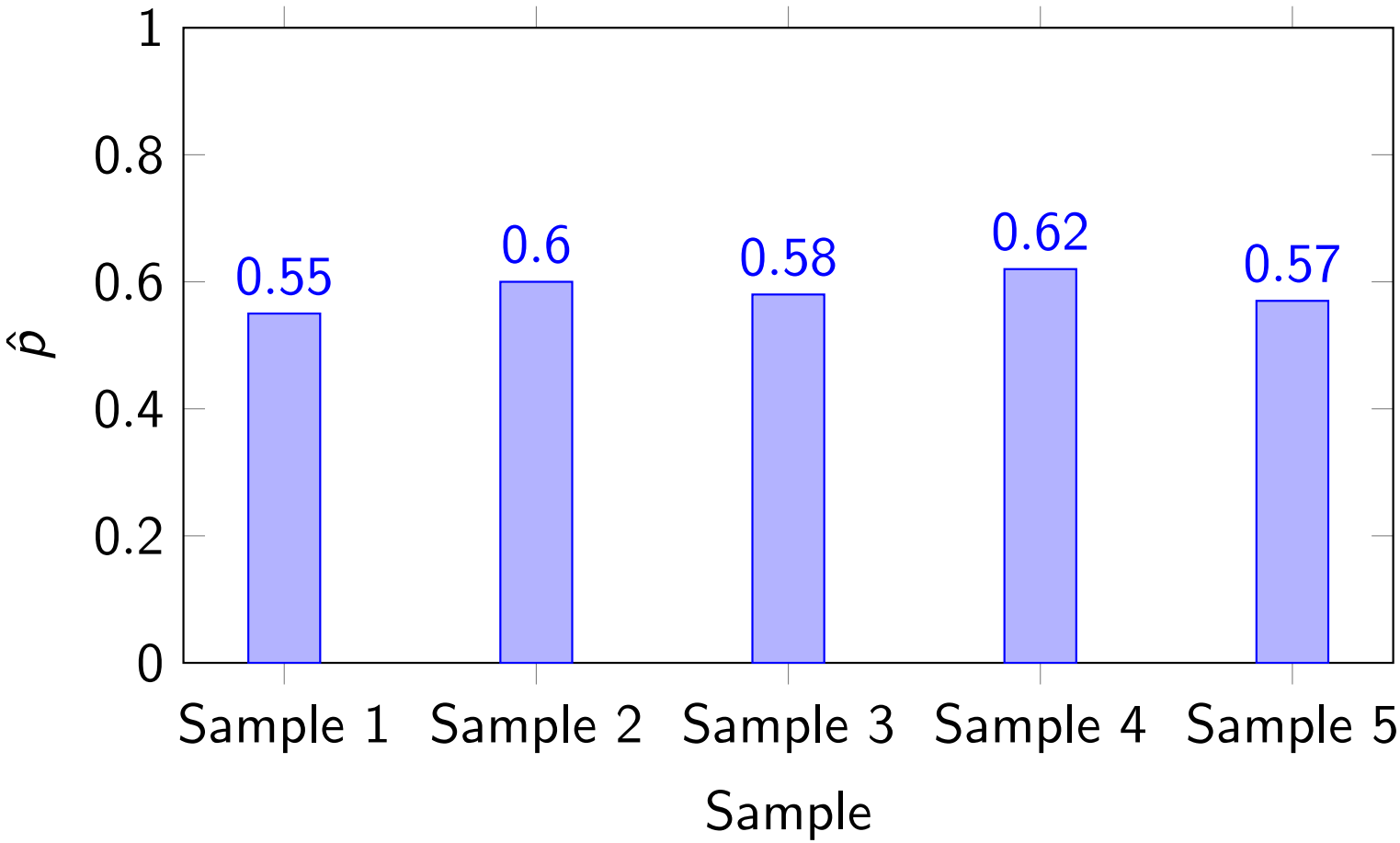

Figure: Variability in Sample Proportions ($\hat{p}$)

Sample Proportion

Let:

$\{X_i\}_{i=1}^N$ be a sample collected via simple random sampling

$$X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Note that any X_k and X_j:

- Are identically distributed. So, $E[X_k] = E[X_j] = p$, and $V[X_k] = V[X_j] = p(1-p)$.
- Are independent. So, this implies the following variance of a linear combination of two values, $V((X_k + X_j)/2) = (1/4) \cdot V(X_k + X_j) = (1/4) \cdot [p(1-p) + p(1-p)] = \frac{p(1-p)}{2}$

Expectation and Variance

Population mean:

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = p$$

Variance:

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)$$

$$= \frac{1}{n^2}\cdot n \cdot p(1-p) = \frac{p(1-p)}{n}$$

Sample Variability and Sampling Distribution

- **Repeated Sampling:** Each sample yields a different $\hat{p}$.
- **Objective:** Quantify the variability of $\hat{p}$.

$$\text{Variance of } \hat{p} = \frac{p(1-p)}{n}$$

$$\text{Standard Error (SE)} = \sqrt{\frac{p(1-p)}{n}}$$

- **Sampling Distribution:** If the sample size is large, we can use the CLT to approximate the sampling distribution of $\hat{p}$:

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Confidence Intervals

What is a Confidence Interval?

- A confidence interval (CI) provides a range of values within which we are fairly certain the true population parameter (e.g., a proportion or mean) lies.
- CIs give us an idea of how reliable our estimate is, based on our sample data.
- They quantify the uncertainty around our estimate, offering a range likely to contain the true value.

Confidence Interval Formula

- For an unknown population proportion p, in large samples the sample proportion $\hat{p} = \frac{S_n}{n}$ is approximately normally distributed:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx Z \sim N(0, 1)$$

- To calculate a 95% confidence interval, we utilize:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

Calculating the Confidence Interval Range

- Translating the standardization back to our interval:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P(-1.96 \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96) \approx 0.95$$

$$P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95$$

- We are 95% confident that the true population parameter p lies within this calculated interval.

Logic Behind the Confidence Interval Derivation

- We start by making a reasonable assumption about the sampling distribution for $\hat{p}$.
- Then we standardize $\hat{p}$ into a Z-score, and estimate the variability due to sampling using the standard normal.
- When we say $P(-1.96 \leq Z \leq 1.96) = 0.95$, we mean that 95% of possible outcomes will fall within this ± 1.96 range under the normal distribution.
- Then, by using the properties of the normal distribution, we define a range within which the true proportion p likely lies.

The Anatomy of a Confidence Interval

- Is a procedure that depends on realized values for a random variables, in this case, $\hat{p}$.
- Thus, it is subject to sample variability.
- Note that for each given confidence interval, either the true proportion is or isn't contained. The true proportion is a fixed quantity.
- We ask, what percentage of the time will the confidence interval capture the true proportion? That is the confidence level, or one minus alpha $(1 - \alpha)$. Alpha is the significance level.
- Over the long run (note this thought exercise is informed by the sampling distribution of $\hat{p}$), we expect that p will be included $(1 - \alpha)\%$ of the time.

Intuition Behind Confidence Intervals

Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 95%.

Estimating the Standard Error (SE)

- The standard error for the proportion is given by:

$$\text{SE} = \sqrt{\frac{p(1-p)}{n}}$$

- Since p is unknown, by the plugin principle, we use the sample proportion $\hat{p}$ to estimate it:

$$\text{SE} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Margin of Error (MOE)

- The margin of error (MOE) is a statistic expressing the amount of random sampling error in the results of a survey.
- The margin of error is found by multiplying the SE by the critical value (1.96 for a 95% confidence level):

$$\text{MOE} = 1.96 \times \text{SE}(\hat{p})$$

- For a general confidence interval level of $(1 - \alpha)$:

$$\text{MOE} = z_{1-\frac{\alpha}{2}} \times \text{SE}(\hat{p})$$

where $z_{1-\frac{\alpha}{2}}$ is the critical value from the standard normal distribution.

Constructing the Confidence Interval

- Finally, the confidence interval is given by:

$$[\hat{p} - \text{MOE}, \hat{p} + \text{MOE}]$$

- For a 95% CI:

$$\text{MOE} = 1.96 \times \text{SE}(\hat{p}) = 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- This interval provides an estimated range that, with 95% confidence, contains the true population proportion p.

General Formula for Confidence Intervals for p

- Invoking the CLT:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- Thus, the Z-score standardized form:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

- To construct a confidence interval with confidence level $1-\alpha$, we use the critical value $z_{1-\alpha/2}$ such that:

$$P\left(Z < z_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}$$

- The general confidence interval for p becomes:

$$\hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Steps to Calculate a 99% Confidence Interval

1. **Determine the Confidence Level and Critical Value:**
   - For a 99% confidence interval, the critical value from the normal distribution is approximately $Z = 2.576$.
2. **Estimate the Standard Error (SE):**
   - Use the sample proportion $\hat{p}$ to estimate SE:

$$\text{SE} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

3. **Calculate the Margin of Error (MOE):**
   - Multiply the SE by the critical value:

$$\text{MOE} = 2.576 \times \text{SE}$$

4. **Construct the Confidence Interval:**
   - Add and subtract the MOE from the sample proportion $\hat{p}$:

$$[\hat{p} - \text{MOE},\ \hat{p} + \text{MOE}]$$

   - This interval provides a 99% confidence range for the true population proportion p.

Applied Example: Computing 95% and 99% Confidence Intervals

Problem Statement:

- A public health survey finds that out of a sample of 400 people, 120 are vaccinated.
- We wish to calculate the 95% and 99% confidence intervals for the true proportion of vaccinated individuals in the population.

Given Data:

- Sample size $(n) = 400$
- Sample proportion $(\hat{p}) = \frac{120}{400} = 0.300$

Step 1: Calculate Standard Error (SE)

- **Formula:** $\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Calculation:**

$$\text{SE} = \sqrt{\frac{0.300 \times (1 - 0.300)}{400}} = \sqrt{\frac{0.300 \times 0.700}{400}} = \sqrt{\frac{0.210}{400}}$$
$$= \sqrt{0.000525}$$
$$\approx 0.023$$

Step 2: Calculate 95% Confidence Interval

- **Critical Value:** For a 95% confidence level, $Z_{1-0.05/2} = Z_{0.975} = 1.96$
- **Margin of Error (MOE):**

$$\text{MOE} = 1.96 \times \text{SE} = 1.96 \times 0.023 \approx 0.045$$

- **95% CI Calculation:**

$$\text{CI} = \hat{p} \pm \text{MOE} = 0.300 \pm 0.045$$

$$95\%\ \text{CI} = [0.255, 0.345]$$

Step 3: Calculate 99% Confidence Interval

- **Critical Value:** For a 99% confidence level (i.e., $\alpha = 0.01$), $Z_{1-0.01/2} = Z_{0.995} = 2.576$
- **Margin of Error (MOE):**

$$\text{MOE} = 2.576 \times \text{SE} = 2.576 \times 0.023 \approx 0.059$$

- **99% CI Calculation:**

$$\text{CI} = \hat{p} \pm \text{MOE} = 0.300 \pm 0.059$$

$$99\%\ \text{CI} = [0.241, 0.359]$$

Interpretation of 95% and 99% Confidence Intervals

- **95% CI (0.255, 0.345):**
  - We are 95% confident that the true proportion of vaccinated individuals is between 25.5% and 34.5%.
- **99% CI (0.241, 0.359):**
  - We are 99% confident that the true proportion of vaccinated individuals is between 24.1% and 35.9%.
- **Observations:**
  - The 99% CI is wider than the 95% CI, reflecting greater confidence and a broader range for the estimate.

Which Percent of Future Sample Proportions Will Fall in Our CI?

- A 95% confidence interval captures the **true parameter** p in 95% of repeated samples.
- But the interval is built around one observed sample proportion $\hat{p}_{\text{orig}}$.
- **Future sample proportions** $\hat{p}_{\text{new}}$ follow a distribution centered at p, not at $\hat{p}_{\text{orig}}$.
- Therefore, the long-run probability that future $\hat{p}_{\text{new}}$ falls within the **fixed CI** from the past will be most likely lower than 95%.

But note this is a **moot question**. We really care about p, not $\hat{p}$!

Hypothesis Testing

Introduction to Hypothesis Testing

- **Why Test Hypotheses?**
  - Often, we want to know if a certain belief or assumption about a population is likely to be true based on our sample data.
  - For example, we might wonder, "Is the vaccination rate really 30% in the general population, or is it different?"
- **Hypothesis Testing:** A systematic way to check if our data supports or refutes our initial belief.
  - Imagine having a statement about a population—our hypothesis—and then using data to evaluate if there's enough evidence to challenge that statement.

How Hypothesis Testing Works

- **Formulating Two Opposing Claims:**
  - We start with two possible claims about a population measure, such as a proportion.
  - One claim (the *null hypothesis*) represents a baseline assumption, often suggesting "no change" or "no effect."
  - The other claim (the *alternative hypothesis*) suggests there's a meaningful difference from the null.
- **Gathering Evidence:**
  - Using sample data, we evaluate if there's enough evidence to support or refute our baseline assumption.
  - Just as in a courtroom, we start with a presumption (the null hypothesis is true) and only reject it if the evidence is strong.
- **Making a Decision:**
  - If our sample data aligns well with the null hypothesis, we "fail to reject" it.
  - If the data strongly contradicts the null hypothesis, we "reject" it in favor of the alternative hypothesis.

Introducing the Test Statistic

- **What is a Test Statistic?**
  - A test statistic is a number we calculate from our sample data to help us decide if our sample provides enough evidence to challenge the null hypothesis.
  - Think of it as a "score" that tells us how far our sample result is from what we would expect if the null hypothesis were true.
- **The Formula for the Test Statistic (for Proportions):**

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- **Defining Each Term:**
  - $\hat{p}$: The **sample proportion**.
  - p_0: The **null hypothesized proportion**, or the value of the proportion we are assuming is true under the null hypothesis.
  - n: The **sample size**, or the number of individuals in our sample.
  - $\sqrt{\frac{p_0(1-p_0)}{n}}$: The **standard error (SE)** of $\hat{p}$ assuming the null hypothesis.

Understanding the Test Statistic and its Interpretation

- **What Units Does the Test Statistic Use?**
  - The test statistic Z is measured in **standard error units**.
  - It represents how many standard errors the sample proportion $\hat{p}$ is away from the null hypothesis proportion p_0.
- **Why Large Absolute Values of Z Suggest Evidence Against the Null:**
  - If Z is large (either positive or negative), it means $\hat{p}$ is far from p_0 in terms of *expected variability under the assumption of the null hypothesis*, which may indicate that the null hypothesis is unlikely.

Formal Setup of a Hypothesis Test

- **Defining Hypotheses:**
  - We write the **null hypothesis** as: $H_0 : p = p_0$
  - The **alternative hypothesis** represents what we want to test against H_0. For example:
    - $H_a : p \neq p_0$ (two-tailed)
    - $H_a : p > p_0$ (one-tailed, right)
    - $H_a : p < p_0$ (one-tailed, left)
- **Significance Level (α):**
  - α is the threshold probability for deciding when to reject H_0.
  - Common values for α are 0.05 or 0.01, indicating a 5
  - If the probability of observing our test statistic (or more extreme) under H_0 is less than α, we reject H_0.
  - α is a standard, carefully chosen rule that guides us on when to be confident enough to reject H_0 based on how unlikely our observed data is under the null hypothesis.

Intuition of the Significance Level

- **Understanding α as a Tolerance for Error:**
  - The significance level α defines how much evidence we need to reject H_0.
  - It represents our tolerance for being wrong—a boundary for the probability of making a Type I error (rejecting H_0 when it's actually true).
  - Typical values (e.g., $\alpha = 0.05$) mean we accept up to a 5% risk of mistakenly rejecting H_0.
- **The Rejection Region:**
  - The rejection region is determined by α and lies at the "extreme ends" of the distribution under H_0.
  - If our test statistic falls in this region, it suggests that our observed result is too rare under H_0 for us to retain it with confidence.
  - Thus, if our result is in the rejection region, we are willing to reject H_0 because it fits our set threshold for "unusual" results.

Example: Hypothesis Test for Voting Preference

- **Context:** Pew Research surveyed 1,000 participants on voting preference.
  - 52% of the sample say they will vote for Trump.
  - We want to test if the true proportion who will vote for Harris is 51% (i.e., the true proportion of Trump voters is 49%).
- **Hypotheses:**
  - Null hypothesis $H_0 : p = 0.49$
  - Alternative hypothesis $H_a : p \neq 0.49$ (two-tailed test)
- **Significance Level:** Set $\alpha = 0.05$

Solution: Test Statistic and Conclusion

- **Sample Proportion:** $\hat{p} = 0.52$
- **Standard Error (SE):**

$$\text{SE} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.49 \times (1-0.49)}{1000}} \approx 0.0158$$

- **Test Statistic:**

$$Z = \frac{\hat{p} - p_0}{\text{SE}} = \frac{0.52 - 0.49}{0.0158} \approx 1.898$$

- **Decision:**
  - Since $Z = 1.898$ is within the range of $[-1.96, 1.96]$, we **fail to reject** H_0 at $\alpha = 0.05$.
  - Conclusion: There is not enough evidence to suggest that the true proportion differs from 49%.

Rejection Region, Alpha, and Z Test Statistic

- **Defining the Rejection Region:**
  - The **rejection region** is determined by the significance level (α) and represents the values of the test statistic for which we reject H_0.
  - It depends on whether the test is **one-sided** or **two-sided**.
- **Critical Values and Non-Rejection Region:**
  - For a **two-sided test** at significance level α:

$$\text{Non-Rejection Region: } -Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$$

  - For a **one-sided test** (right-tailed) at significance level α:

$$\text{Non-Rejection Region: } Z \leq Z_{\alpha}$$

  - For a **one-sided test** (left-tailed):

$$\text{Non-Rejection Region: } Z \geq -Z_{\alpha}$$

  - The critical value Z_{α} is found from standard normal tables corresponding to the chosen α.
- **Decision Rule:**
  - If the calculated Z falls **inside** the rejection region, we **reject** H_0.

Understanding the P-value

- **Definition of P-value:**
  - The P-value is the probability, under the null hypothesis H_0, of obtaining a test statistic as extreme as, or more extreme than, the observed value.
- **Relation to Z Test Statistic:**
  - For a **two-sided test**:

$$\text{P-value} = 2 \times P(Z \geq |Z_{\text{obs}}|)$$

  - For a **one-sided test** (right-tailed):

$$\text{P-value} = P(Z \geq Z_{\text{obs}})$$

  - For a **one-sided test** (left-tailed):

$$\text{P-value} = P(Z \leq Z_{\text{obs}})$$

- **Interpreting the P-value:**
  - A small P-value indicates strong evidence against H_0.
  - A large P-value suggests that the observed data is consistent with H_0.

P-value and Different Significance Levels

- **Assessing Significance with P-value:**
  - The P-value allows us to determine at which significance levels H_0 would be rejected.
  - By comparing the P-value to various α levels, we can see the minimum α for which we would reject H_0.
- **Decision Making:**
  - If P-value $\leq \alpha$, we **reject** H_0.
  - If P-value $> \alpha$, we **fail to reject** H_0.
- **Example:**
  - If P-value $= 0.03$, H_0 would be rejected at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Example: Computing P-value for a One-Sided Test

Problem Statement:

- A political poll indicates that 52% of a sample of 1,000 voters support Candidate A.
- We wish to test if there is evidence that the true proportion supporting Candidate A is greater than 50%.

Hypotheses:

- Null hypothesis $H_0 : p = 0.50$
- Alternative hypothesis $H_a : p > 0.50$ (one-sided test)

Calculations:

- Sample proportion $\hat{p} = 0.52$
- Standard Error SE $= \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.50 \times 0.50}{1000}} \approx 0.0158$
- Test Statistic $Z = \frac{\hat{p}-p_0}{\text{SE}} = \frac{0.52-0.50}{0.0158} \approx 1.265$
- P-value $= P(Z \geq 1.265) = 1 - \Phi(1.265) \approx 0.103$

Conclusion:

- At $\alpha = 0.05$, since P-value $= 0.103 > 0.05$, we **fail to reject** H_0.
- There is insufficient evidence to reject the hypothesis that $p_0 = 0.50$